IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of: Martin M. Deneroff et al.
Title: NETWORK TOPOLOGY FOR A SCALABLE MULTIPROCESSOR SYSTEM
Attorney Docket No.: 499.038US1

PATENT APPLICATION TRANSMITTAL

BOX PATENT APPLICATION
Assistant Commissioner for Patents
Washington, D.C. 20231

We are transmitting herewith the following attached items and information (as indicated with an "X"):

X Utility Patent Application under 37 CFR § 1.53(b) comprising:
- X Specification (19 pgs, including claims numbered 1 through 6 and a 1 page Abstract).
- X Formal Drawing(s) (17 sheets).
- X Unsigned Combined Declaration and Power of Attorney (4 pgs).

X Return postcard.

The filing fee (NOT ENCLOSED) will be calculated as follows:

	No. Filed	No. Extra	Rate	Fee
TOTAL CLAIMS	6 - 20 =	0	x 18 =	**$0.00**
INDEPENDENT CLAIMS	2 - 3 =	0	x 78 =	**$0.00**
MULTIPLE DEPENDENT CLAIMS PRESENTED				**$0.00**
BASIC FEE				**$760.00**
TOTAL				**$760.00**

THE FILING FEE WILL BE PAID UPON RECEIPT OF THE NOTICE TO FILE MISSING PARTS.

SCHWEGMAN, LUNDBERG, WOESSNER & KLUTH, P.A.
P.O. Box 2938, Minneapolis, MN 55402 (612-373-6900)

Customer Number 21186

By: John M. Dahl
Atty: John M. Dahl
Reg. No. P-44,639

"Express Mail" mailing label number: EL273420485US Date of Deposit: September 29, 1999
I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Box Patent Application, Washington, D.C. 20231.

By: Chris Hammond Signature: Chris Hammond

NETWORK TOPOLOGY FOR A SCALABLE MULTIPROCESSOR SYSTEM

Field of the Invention

The present invention relates generally to the field of high-speed digital data processing systems, and more particularly, to interconnection topologies for interconnecting processing element nodes in multiprocessor computer systems.

Background of the Invention

Multiprocessor computer systems comprise a number of processing element nodes connected together by an interconnect network. Each processing element node includes at least one processing element. The interconnect network transmits packets of information or messages between processing element nodes. Multiprocessor computer systems having up to hundreds or thousands of processing element nodes are typically referred to as massively parallel processing (MPP) systems. In a typical multiprocessor MPP system, every processing element can directly address all of memory, including the memory of another (remote) processing element, without involving the processor at that processing element. Instead of treating processing element-to-remote-memory communications as an I/O operation, reads or writes to another processing element's memory are accomplished in the same manner as reads or writes to the local memory. In such multiprocessor MPP systems, the infrastructure that supports communications among the various processors greatly affects the performance of the MPP system because of the level of communications required among processors.

Several different topologies have been proposed to interconnect the various processors in such MPP systems, such as rings, stars, meshes, hypercubes, and torus topologies. For example, in a conventional hypercube network, a plurality of microprocessors are arranged in an n-dimensional cube where the number of nodes k in the network is equal to 2^n. In this network, each node is connected to each other

node via a plurality of communications paths. The network diameter, the longest communications path from any one node on the network to any other node, is *n*-links.

Regardless of the topology chosen, one disadvantage of current multiprocessor systems, and in particular MPP systems, is that in order to expand the system, a significant amount of reconfiguration is required. The reconfiguration often involves removing and replacing cables which is very time consuming. Also, as systems increase the number of processors, the number of physical connections required to support the system increases significantly which increases the complexity of the system.

Therefore, it is desired that systems could be easily scaled to increase the number of processors with minimal disruption to the original system configuration.

Summary of the Invention

The present invention provides a system and method for interconnecting a plurality of processing element nodes within a scalable multiprocessor system. Each processing element node includes at least one processor and memory. A scalable interconnect network includes physical communication links interconnecting the processing element nodes in a cluster. A first set of routers in the scalable interconnect network route messages between the plurality of processing element nodes. One or more metarouters in the scalable interconnect network route messages between the first set of routers so that each one of the routers in a first cluster is connected to all other clusters through one or more metarouters.

Brief Description of the Drawings

FIG. 1 is block diagram of a multiprocessor computer system.

FIG. 2 is a block diagram of one embodiment of the interface between a scalable interconnect network and four processing element nodes.

FIG. 3 is a model of a two dimensional (2D) hypercube topology multiprocessor computer system.

FIG. 4 is a model of a three dimensional (3D) hypercube topology multiprocessor computer system.

FIG. 5 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having 129 processors to 160 processors.

FIG. 6 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having 161 processors to 192 processors.

FIG. 7 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having 193 processors to 224 processors.

FIG. 8 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having 225 processors to 256 processors.

FIG. 9 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 288 processors.

FIG. 10 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 320 processors.

FIG. 11 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 352 processors.

FIG. 12 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 384 processors.

FIG. 13 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 416 processors.

FIG. 14 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 448 processors.

FIG. 15 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 480 processors.

FIG. 16 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 512 processors.

FIG. 17 illustrates an example embodiment of a multiprocessor computer system having 512 processors.

Description of the Preferred Embodiments

In the following detailed description of the preferred embodiments, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration specific embodiments in which the invention may be practiced. It is to be understood that other embodiments may be utilized and structural or logical changes may be made without departing from the scope of the present invention. The following detailed description, therefore, is not to be taken in a limiting sense, and the scope of the present invention is defined by the appended claims.

System Overview. A representative multiprocessor computer system according to the present invention is indicated generally at 20 in Figure 1. As indicated in Figure 1, multiprocessor computer system 20 includes up to n nodes, such as indicated by a first node 22, a second node 24, and an nth node 26. The nodes are interconnected by a scalable interconnect network 28, which permits multiprocessor computer systems 20 to be scaled from desk side systems to very large supercomputer configurations.

As illustrated in detail for first node 22, each node in multiprocessor computer system 20 includes at least one processor, such as a first processor 30 and a second processor 32 for node 22. An interface circuit 34 interfaces with scalable interconnect network 28 and communicates with a memory and directory 36 and an input/output subsystem 38.

Although the multiprocessor computer system 20 illustrated in FIG. 1 provides one example environment to implement the below-described network topology according to the present invention, the present invention is in no way limited to this particular application environment. In fact, many alternative environments using alternative node and interface circuit configurations can be

utilized. To a large extent, the network topology according to the present invention, as implemented in scalable interconnect network 28, is independent of the complexity of the nodes, such as nodes 22, 24, and 26, interconnected by that topology.

FIG. 2 illustrates, in block diagram form, one embodiment of the interface between a scalable interconnect network 128 and four nodes 122, 123, 124, and 125. In this embodiment, scalable interconnect network 128 includes one or more routers, such as the example router 150 shown in FIG. 2. Router port 163 communicates with a first node 122. As shown in the expanded view of the first node 122, router port 163 communicates with the first node 122 via interface chip 134. In the first node 122, interface chip 134 communicates with processors 130, 131, 132, and 133. Similarly, router port 164 communicates with a second node 123, router port 165 communicates with a third node 124, and router port 166 communicates with a fourth node 125. Router ports 152, 154, 156, and 158 communicate with other routers in the scalable interconnect network 128 as further described below. In the example embodiment illustrated in FIG. 2, one router 150 communicates directly with up to sixteen processors and up to four other routers in the scalable interconnect network 128.

The scalable interconnect network 128 of the present invention employs a first set of routers (such as router 150 of FIG. 2) for routing messages between a plurality of processing element nodes (such as nodes 122, 123, 124, and 125 in FIG. 2). In one embodiment of the present invention, four of the eight ports of each one of the first set of routers are dedicated to connecting from the routers to four separate nodes, such as indicated in FIG. 2. As further described below, the remaining ports are connected to the first set of routers for routing messages between the nodes and a second set of routers (referred to herein as "metarouters") for routing messages between the first set of routers. In one embodiment, the metarouters shown in FIGS. 5-16 below are multiport routers having at least four ports. One of skill in the art will recognize that a single multiport router or any

combination of multiport routers can be used in the configurations shown in FIGS. 5-16.

As will be better understood by the following discussion, the topology of a scalable interconnect network according to the present invention, is easily scaleable, has increased resiliency and allows system upgrades/expansion to be performed with minimal disruption to the system.

Example Network Topologies. A grouping of a plurality of processing element nodes, a plurality of physical interconnections (such as cables) for connecting the plurality of processing element nodes, and a first set of routers for routing messages between a plurality of processing element nodes is referred to herein as a "cluster." In a multiprocessing system of the present invention, each cluster is connected to each one of the other clusters using one or more metarouters. Thus, each one of the clusters can communicate directly with another one of the clusters without having to communicate through a third cluster.

In one embodiment, the processing element nodes and first set of routers form one or more two-dimensional hypercube systems. An example two dimensional (2D) hypercube topology multiprocessor computer system is shown in FIG. 3. In FIG. 3, the four routers 150 are numbered 0 through 3 and are interconnected as a 2D hypercube. In one embodiment, the 2D hypercube comprises up to sixty-four processors because each of the four routers communicates with up to sixteen processors as shown in FIG. 2. In the following detailed description clusters are configured in a two-dimensional hypercube for illustrative purposes only. The invention is not limited to arranging the first set of routers in a two-dimensional hypercube. Alternate embodiments in which the first set of routers are grouped in different topologies are contemplated as within the scope of the invention.

In one embodiment, a multiprocessor computer system is constructed with up to 128 processors without the use of metarouters. Such an embodiment is shown in Fig. 4. in which a first cluster 400 and a second cluster 402 are each arranged as

two-dimensional hypercubes. As shown in FIG. 4, the routers of a first cluster 400 and a second cluster 402 are interconnected to form a three-dimensional (3D) hypercube. The 3D hypercube is comprised of eight routers 150 numbered 0 through 7. In an example embodiment, each one of the routers is an eight port router connected to four processing element nodes such as the router shown in FIG. 2. For processing element nodes having four processors each, the 3D hypercube interconnects up to 128 processors.

Larger scale system configurations and the corresponding logical topologies for example systems having 128 or more processors are described below. The larger scale systems employ a second set of routers (referred to herein as "metarouters") to interconnect the clusters. The novel network topologies of the present invention allow two clusters to communicate directly without having to route messages through a third cluster. Rather than routing messages through a third cluster, the routers in each cluster are connected to all other clusters through the metarouters.

As shown in FIGS. 5-16, each router is either connected directly to a router of another cluster or is connected to a router of another cluster using one or more metarouters. In the example embodiments shown in FIGS. 5-16, the connections from a single router in one cluster to routers in all other clusters are represented by particular line styles. The dashed lines in FIGS. 5-16 represent the connections from a first router in a first cluster to a single router in each one of the other clusters. The solid lines in FIGS. 5-16 represent the connections from a second router in the first cluster to a second router in each one of the other clusters. The dotted lines in FIGS. 5-16 represent the connections from a third router in the first cluster to a third router in each one of the other clusters. The dash/dot lines in FIGS. 5-16 represent the connection from a fourth router in a first cluster to a fourth router in each one of the other clusters.

FIG. 5 illustrates an example embodiment of a logical topology used for systems having 129 processors to 160 processors. The logical topology comprises three clusters 502, 504, 506 interconnected with two metarouters 508, 510. In an

alternate embodiment, a single multiport metarouter is used instead of the two metarouters 508, 510. In the example embodiment, a first cluster 502 comprises sixty-four processors and the second cluster 504 also comprises sixty-four processors. The third cluster 506 can include any number of processors from one to thirty-two making a total of 129 to 160 processors in the multiprocessor computer system shown in FIG. 5.

A router in each one of the three clusters 502, 504, 506 is connected to another router in each one of the other three clusters 502, 504, 506 through the metarouters 508, 510. Dashed lines represent the connection between a first router in a first cluster and a single router in each one of the other clusters. The first router 512 in a first cluster 502 is connected to a first router 520 in a second cluster 504 through a metarouter 508. The first router 512 in the first cluster 502 is also connected to a first router 530 in a third cluster 506 through the metarouter 508.

Likewise, the solid lines represent the connection between a second router in the first cluster with a second router in each one of the other clusters. The second router 514 in the first cluster 502 is connected to a second router 522 in the second cluster 504 through a metarouter 510. The second router 514 in the first cluster 502 is also connected to a second router 532 in third cluster 506 through metarouter 510.

Similarly, the dotted lines represent the connection between a third router in the first cluster with a third router in each one of the other clusters and dash/dot lines in the diagrams represent the connection between a fourth router in the first cluster with a fourth router in each one of the other clusters. The third router 516 and the fourth router 518 in the first cluster 502 are directly connected to a third router 524 and a fourth router 528 in the second cluster 504. The third cluster 506 lacks a third router and a fourth router in this configuration. Thus, as shown in FIG. 5, each one of the routers in a cluster communicates with a router in each of the other clusters through either a direct connection or through one or more metarouters. This eliminates the dependancy of one cluster to communicate with another cluster through a third cluster and thus creates a more resilient system. The topologies of

the present invention allow clusters to communicate directly with each other without communicating through a third cluster.

FIG. 6 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having 161 processors to 192 processors. The logical topology comprises three clusters 602, 604, 606 interconnected with four metarouters 608, 609, 610, and 611. In an alternate embodiment, two multiport metarouters are used in place of the four metarouters 608, 609, 610 and 611. In the example embodiment shown in FIG. 6, a first cluster 602 comprises sixty-four processors and a second cluster 604 also comprises sixty-four processors. A third cluster 606 can include any number of processors from thirty-three processors to sixty-four processors for a total of 161 to 192 processors in the multiprocessor computer system shown in FIG. 6.

A router in each one of the three clusters 602, 604, 606 is connected to another router in each one of the other three clusters 602, 604, 606 through the metarouters 608, 609, 610 and 611. The first router 612 in a first cluster 602 is connected to a first router 620 in a second cluster 604 through a metarouter 608. The first router 612 in the first cluster 602 is also connected to a first router 630 in a third cluster 606 through the metarouter 608. Connections between the first router 612 in the first cluster 602 and each one of the other clusters are represented in FIG. 6 with dashed lines.

Likewise, the second router 614 in the first cluster 602 is connected to a second router 622 in the second cluster 604 through a metarouter 610. The second router 614 in the first cluster 602 is also connected to a second router 632 in third cluster 606 through metarouter 610. Connections between the second router 614 in the first cluster 602 and each one of the other clusters are represented in FIG. 6 with solid lines. Similarly, the dotted lines in FIG. 6 represent the connections between a third router 616 in the first cluster 602 and each one of the other clusters. The third router 616 in the first cluster 602 is connected to a third router 624 in the second cluster 604 through metarouter 609. The third router 616 in the first cluster 602 is

also connected to a third router 634 in the third cluster 606 through metarouter 609. Additionally, the dash/dot lines in FIG. 6 represent the connections between a fourth router 618 in the first cluster 602 and each one of the other clusters. The fourth router 618 in the first cluster 602 is connected to a fourth router 628 in the second cluster 604 through metarouter 611. The fourth router 618 in the first cluster 602 is also connected to a fourth router 636 in the third cluster 606 through metarouter 611. Thus, as shown in FIG. 6, each one of the routers in a cluster communicates with a router in each of the other clusters through metarouters.

FIG. 7 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having 193 processors to 224 processors. The logical topology comprises four clusters 702, 704, 706, 707 interconnected with four metarouters 708, 709, 710, and 711. In an alternate embodiment, two multiport metarouters are used in place of the four metarouters 708, 709, 710, and 711. In the example embodiment shown in FIG. 7, a first cluster 702, a second cluster 704, and a third cluster 706 each comprise sixty-four processors. A fourth cluster 707 includes any number of processors from one processor up to thirty-two processors for a total of 193 to 224 processors in the multiprocessor computer system shown in FIG. 7.

A router in each one of the four clusters 702, 704, 706, 707 is connected to another router in each one of the other four clusters 702, 704, 706, 707 through one of the metarouters 708, 709, 710 and 711. Connections between the first router 712 in the first cluster 702 and each one of the other clusters are represented in FIG. 7 with dashed lines. The first router 712 in a first cluster 702 is connected to a first router 720 in a second cluster 704 through a metarouter 708. The first router 712 in the first cluster 702 is also connected to a first router 730 in a third cluster 706 through the metarouter 708. In addition, as shown in FIG. 7, the first router 712 in the first cluster 702 is connected to a first router 738 in a fourth cluster 707 through the metarouter 708.

Connections between the second router 714 in the first cluster 702 and each one of the other clusters are represented in FIG. 7 with solid lines. The second router 714 in the first cluster 702 is connected to a second router 722 in the second cluster 704 through a metarouter 710. The second router 714 in the first cluster 702 is also connected to a second router 732 in third cluster 706 through metarouter 710 Additionally, in FIG. 7 the second router 714 in the first cluster 702 is connected to a second router 740 in a fourth cluster 707 through the metarouter 708.

Similarly, the dotted lines in FIG. 7 represent the connections between a third router 716 in the first cluster 702 and each one of the other clusters. The third router 716 in the first cluster 702 is connected to a third router 724 in the second cluster 704 through metarouter 709. The third router 716 in the first cluster 702 is also connected to a third router 734 in the third cluster 706 through metarouter 709. Likewise, the dash/dot lines in FIG. 7 represent the connections between a fourth router 718 in the first cluster 702 and each one of the other clusters. The fourth router 710 in the first cluster 702 is connected to a fourth router 728 in the second cluster 704 through metarouter 711. The fourth router 710 in the first cluster 702 is also connected to a fourth router 736 in the third cluster 706 through metarouter 711.

Thus, as shown in FIG. 7, each one of the routers in a cluster communicates with a router in each of the other clusters through metarouters. Furthermore, the 224 processor system shown in FIG. 7 is easily created by expanding the 192 processor system shown in FIG. 6. First, the multiprocessor system of FIG.7 is formed by adding the fourth cluster 707. Second, the interconnections are added from the metarouter 708 to a first router 738 in the fourth cluster 707. Third, the interconnections are added from the metarouter 710 to the second router 740 in a fourth cluster 707. Thus, a 192 processor system of the present invention is expandable to a 224 processor system without having to recable the first, the second, and the third clusters 702, 704, 706 to expand the configuration.

FIG. 8 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having 225 processors to 256 processors. The logical topology comprises four clusters 802, 804, 806, 807 interconnected with four metarouters 808, 809, 810, and 811. In the example embodiment shown in FIG. 8, a first cluster 802, a second cluster 804, and a third cluster 806 each contain sixty-four processors. In addition, the fourth cluster 807 can include from thirty-three processors up to sixty-four processors for a total of 225 to 256 processors in the multiprocessor computer system shown in FIG. 8.

A router in each one of the four clusters 802, 804, 806, 807 is connected to another router in each one of the other four clusters 802, 804, 806, 807 through one of the metarouters 808, 809, 810 and 811. Connections between the first router 812 in the first cluster 802 and each one of the other clusters are represented in FIG. 8 with dashed lines. The first router 812 in a first cluster 802 is connected to a first router 820 in a second cluster 804 through a metarouter 808. The first router 812 in the first cluster 802 is also connected to a first router 830 in a third cluster 806 through the metarouter 808. The first router 812 in the first cluster 802 is also connected to a first router 838 in a fourth cluster 807 through the metarouter 808.

Connections between the second router 814 in the first cluster 802 and each one of the other clusters are represented in FIG. 8 with solid lines. The second router 814 in the first cluster 802 is connected to a second router 822 in the second cluster 804 through a metarouter 810. The second router 814 in the first cluster 802 is also connected to a second router 832 in third cluster 806 through metarouter 810, and the second router 814 in the first cluster 802 is connected to a second router 840 in a fourth cluster 807 through the metarouter 808.

Similarly, the dotted lines in FIG. 8 represent the connections between a third router 816 in the first cluster 802 and each one of the other clusters. The third router 816 in the first cluster 802 is connected to a third router 824 in the second cluster 804 through metarouter 809. The third router 816 in the first cluster 802 is also connected to a third router 834 in the third cluster 806 through metarouter 809.

Additionally, the 224 processor system of FIG. 7 is expanded in FIG. 8 by adding a processing element node and a third router 842 in the fourth cluster 807 and connecting the third router 816 in the first cluster 802 to the third router 842 in the fourth cluster 807 through metarouter 809.

Likewise, the dash/dot lines in FIG. 8 represent the connections between a fourth router 818 in the first cluster 802 and each one of the other clusters. The fourth router 818 in the first cluster 802 is connected to a fourth router 828 in the second cluster 804 through metarouter 811. The fourth router 818 in the first cluster 802 is also connected to a fourth router 836 in the third cluster 806 through metarouter 811. Additionally, the 224 processor system of FIG. 7 is further expanded in FIG. 8 by adding a processing element node and a fourth router 844 and connecting the fourth router 818 in the first cluster 802 to the fourth router 844 in the fourth cluster 807 through metarouter 811.

Again, as shown in FIG. 8, each one of the routers in a cluster communicates with a router in each of the other clusters through metarouters. Furthermore, the 256 processor system shown in FIG. 8 is easily created by expanding the 224 processor system shown in FIG. 7. Thus, a 224 processor system of the present invention is expandable to a 256 processor system merely by adding processing element nodes, interconnections, and routers to the fourth cluster 807. The expansion of the system does not require reconfiguration or recabling the first, the second, or the third clusters 702, 704, 706 in FIG. 7.

FIGS. 9-16 illustrate example embodiments of multiprocessor systems having greater than 256 processors. The multiprocessor systems shown in FIGS. 9-16 are built by expanding the configuration shown in FIG. 8 comprising up to 256 processors. The 256 processor system shown in FIG. 8 has four clusters configured as 2D hypercubes and interconnected with metarouters. According to the present invention, multiprocessor systems having more than 256 processors are configured by adding clusters to the configuration shown in FIG. 8 and connecting each one of the clusters to the other clusters in the configuration by one or more metarouters.

FIG. 9 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 288 processors. The 288 processor system shown in FIG. 9 is easily created by expanding the 256 processor system shown in FIG. 8. A fifth cluster 962 is added to the multiprocessor system and two additional metarouters 950, 952 are added to the multiprocessor system. Interconnections are added from the metarouter 950 to a first router 970 in the fifth cluster 962, and to another metarouter. Interconnections are also added from the metarouter 952 to a second router 972 in the fifth cluster 962, and to another metarouter. Thus, a 256 processor system of the present invention is easily expandable to a 288 processor system without having to reconfigure the original 256 processor system.

FIG. 10 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 320 processors. The 320 processor system shown in FIG. 10 is easily created by expanding the 288 processor system shown in FIG. 9. A sixth cluster 1064 is added to the multiprocessor system. Interconnections are added from the metarouter 1050 to a first router 1078 in the sixth cluster 1064. Interconnections are also added from the metarouter 1052 to a second router 1080 in the sixth cluster 1064. As shown in FIG. 10, a 288 processor sysem of the present invention is easily expandable to a 320 processor system without having to reconfigure the original 288 processor system.

FIG. 11 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 352 processors. The 352 processor system shown in FIG. 11 is easily created by expanding the 320 processor system shown in FIG. 10. A third router 1174 (and a corresponding processor element node) and a fourth router 1176 (and a corresponding processor element node) are added to the fifth cluster 1162 of the multiprocessor system. Interconnections are added from the metarouter 1154 (which is also added to the multiprocessor system) to the third router 1174 in the fifth cluster 1162. Interconnections are also added from the metarouter 1156 (which is also added to the system) to the fourth router

1176 in the fifth cluster 1162. The new metarouters 1154, 1156 are also interconnected to the previous metarouters. As shown in FIG. 11, a 320 processor sysem of the present invention is easily expandable to a 352 processor system without having to reconfigure the original 320 processor system.

FIG. 12 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 384 processors. The 384 processor system shown in FIG. 12 is easily created by expanding the 352 processor system shown in FIG. 11. A third router 1282 and a fourth router 1284 and the corresponding processing element nodes are added to the sixth cluster 1264 of the multiprocessor system. Interconnections are added from the metarouter 1254 to the third router 1282 in the sixth cluster 1264. Interconnections are also added from the metarouter 1256 to the fourth router 1284 in the sixth cluster 1264. As shown in FIG. 12, a 352 processor sysem of the present invention is easily expandable to a 384 processor system without having to reconfigure the original 352 processor system.

FIG. 13 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 416 processors. The 416 processor system shown in FIG. 13 is easily created by expanding the 384 processor system shown in FIG. 12. A seventh cluster 1386 having a first router 1388 and a second router 1389 are added to the multiprocessor system. Two metarouters 1355, 1357 are also added. Interconnections are added from the metarouter 1355 to the first router 1388 in the seventh cluster 1386 and to one of the metarouters. Interconnections are also added from the metarouter 1357 to the second router 1389 in the seventh cluster 1386 and to one of the other metarouters. As shown in FIG. 13, a 384 processor sysem of the present invention is easily expandable to a 416 processor system without having to reconfigure the original 384 processor system.

FIG. 14 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 448 processors. The 448 processor system shown in FIG. 14 is easily created by expanding the 416 processor system

shown in FIG. 13. An eighth cluster 1487 having a first router 1492 (and corresponding processing element node) and a second router 1493 (and corresponding processing element node) are added to the multiprocessor system. Interconnections are added from the metarouter 1455 to the first router 1492 in the eighth cluster 1487. Interconnections are also added from the metarouter 1457 to the second router 1493 in the eighth cluster 1487. As shown in FIG. 14, a 416 processor sysem of the present invention is easily expandable to a 448 rocessor system without having to reconfigure the original 416 processor system.

FIG. 15 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 480 processors. The 480 processor system shown in FIG. 15 is easily created by expanding the 448 processor system shown in FIG. 14. A third router 1590 and a fourth router 1591 are added the seventh cluster 1586 of the multiprocessor system. Two metarouters are also added 1558, 1559 and interconnections are added from the two metarouters 1558, 1559 to another metarouter. Interconnections are added from the metarouter 1558 to the third router 1590 in the seventh cluster 1586. Interconnections are also added from the metarouter 1559 to the fourth router 1591 in the seventh cluster 1586. As shown in FIG. 15, a 448 processor sysem of the present invention is easily expandable to a 480 processor system without having to reconfigure the original 448 processor system.

FIG. 16 illustrates an example embodiment of a logical topology used for multiprocessor computer systems having up to 512 processors. The 512 processor system shown in FIG. 16 is easily created by expanding the 480 processor system shown in FIG. 15. A third router 1694 (and a corresponding processing element node) and a fourth router 1695 (and a corresponding processing element node) are added the eighth cluster 1687 of the multiprocessor system. Interconnections are added from the metarouter 1658 to the third router 1694 in the eighth cluster 1687. Interconnections are also added from the metarouter 1659 to the fourth router 1695 in the eighth cluster 1687. As shown in FIG. 16, a 480 processor sysem of the

present invention is easily expandable to a 512 processor system without having to reconfigure the original 480 processor system.

FIG. 17 illustrates an example embodiment of a system according to FIG. 16. Each router (19, 27) is connected to four compute nodes (Cbricks). The remaining four ports are connected to other routers. Router 38 provides between routers, routers 27 and other routers 38.

The present invention, as described above, permits smaller systems to be expanded to larger systems with minimal disruption to the original system configuration. Although specific embodiments have been illustrated and described herein for purposes of description of the preferred embodiment, it will be appreciated by those of ordinary skill in the art that a wide variety of alternate and/or equivalent implementations calculated to achieve the same purposes may be substituted for the specific embodiments shown and described without departing from the scope of the present invention. Those with skill in the mechanical, electro-mechanical, electrical, and computer arts will readily appreciate that the present invention may be implemented in a very wide variety of embodiments. This application is intended to cover any adaptations or variations of the preferred embodiments discussed herein. Therefore, it is manifestly intended that this invention be limited only by the claims and the equivalents thereof.

What is claimed is:

1. A massively parallel processing system comprising:

a plurality of processing element nodes;

a scalable interconnection network comprising:

a plurality of physical communication links; and

a plurality of first level routers for interconnecting the plurality of processing element nodes in a cluster; and

one or more metarouters for interconnecting the plurality of first level routers so that each one of the routers in a first cluster is connected to all other clusters through one or more metarouters.

2. The massively parallel processing system of claim 1 wherein each one of the clusters is a two-dimensional hypercube.

3. The massively parallel processing system of claim 1 wherein each one of the metarouters are eight port routers.

4. The massively parallel processing system of claim 1 wherein each one of the metarouters are four port routers.

5. The massively parallel processing system of claim 1 wherein each one of the processing element nodes comprises four processors.

6. A massively parallel processing system comprising:

a plurality of processors;

a first set of routers for interconnecting the plurality of processors as two-dimensional hypercubes; and

a second set of routers for interconnecting the first set of routers wherein the hypercubes remain in tack as the system is expanded.

Abstract of the Disclosure

A system and method for interconnecting a plurality of processing element nodes within a scalable multiprocessor system is provided. Each processing element node includes at least one processor and memory. A scalable interconnect network includes physical communication links interconnecting the processing element nodes in a cluster. A first set of routers in the scalable interconnect network route messages between the plurality of processing element nodes. One or more metarouters in the scalable interconnect network route messages between the first set of routers so that each one of the routers in a first cluster is connected to all other clusters through one or more metarouters.

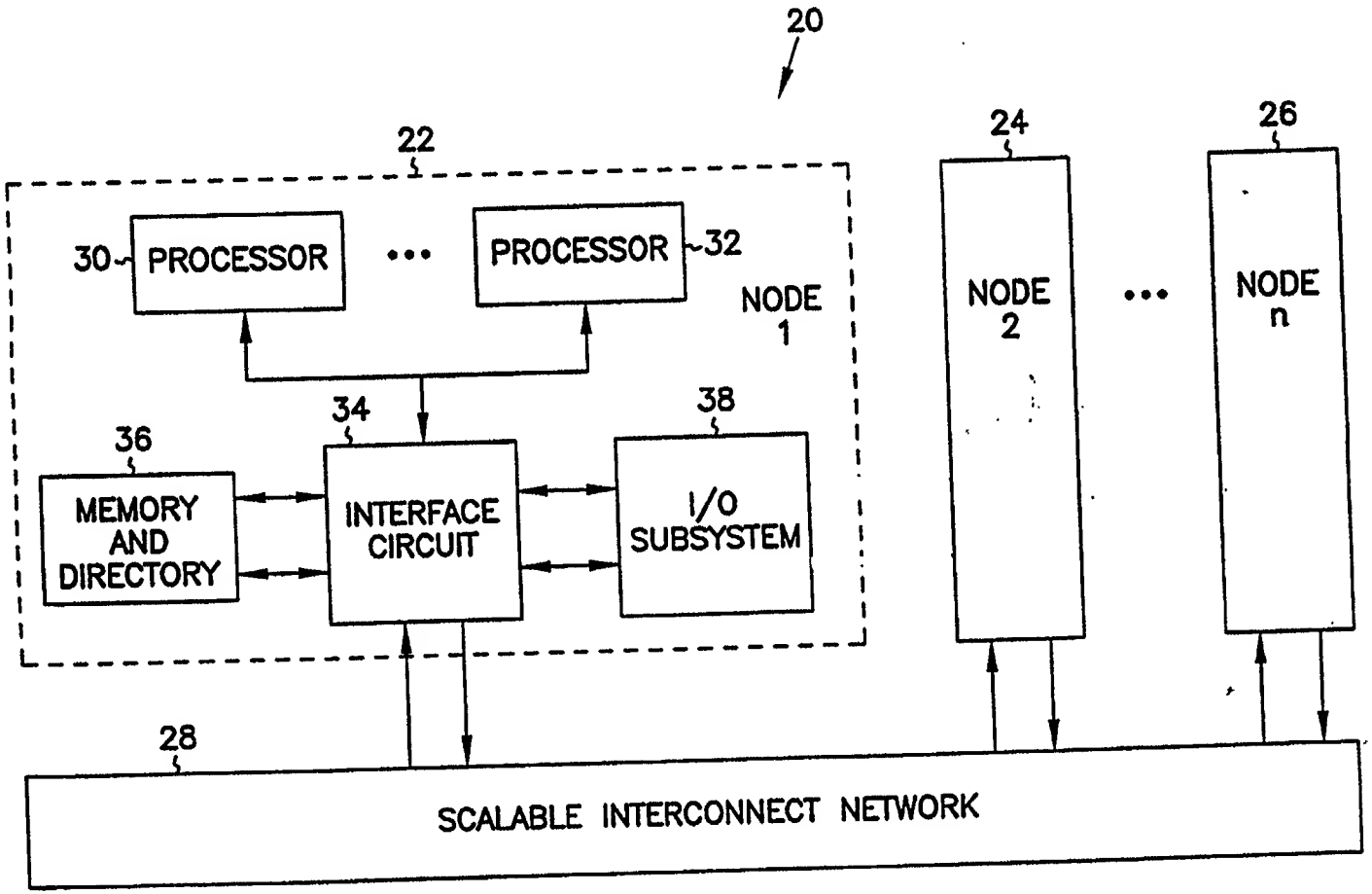

20
22
24
26
30
PROCESSOR
PROCESSOR
32
NODE 1
NODE 2
NODE n
34
36
38
MEMORY AND DIRECTORY
INTERFACE CIRCUIT
I/O SUBSYSTEM
28
SCALABLE INTERCONNECT NETWORK

FIG. 1

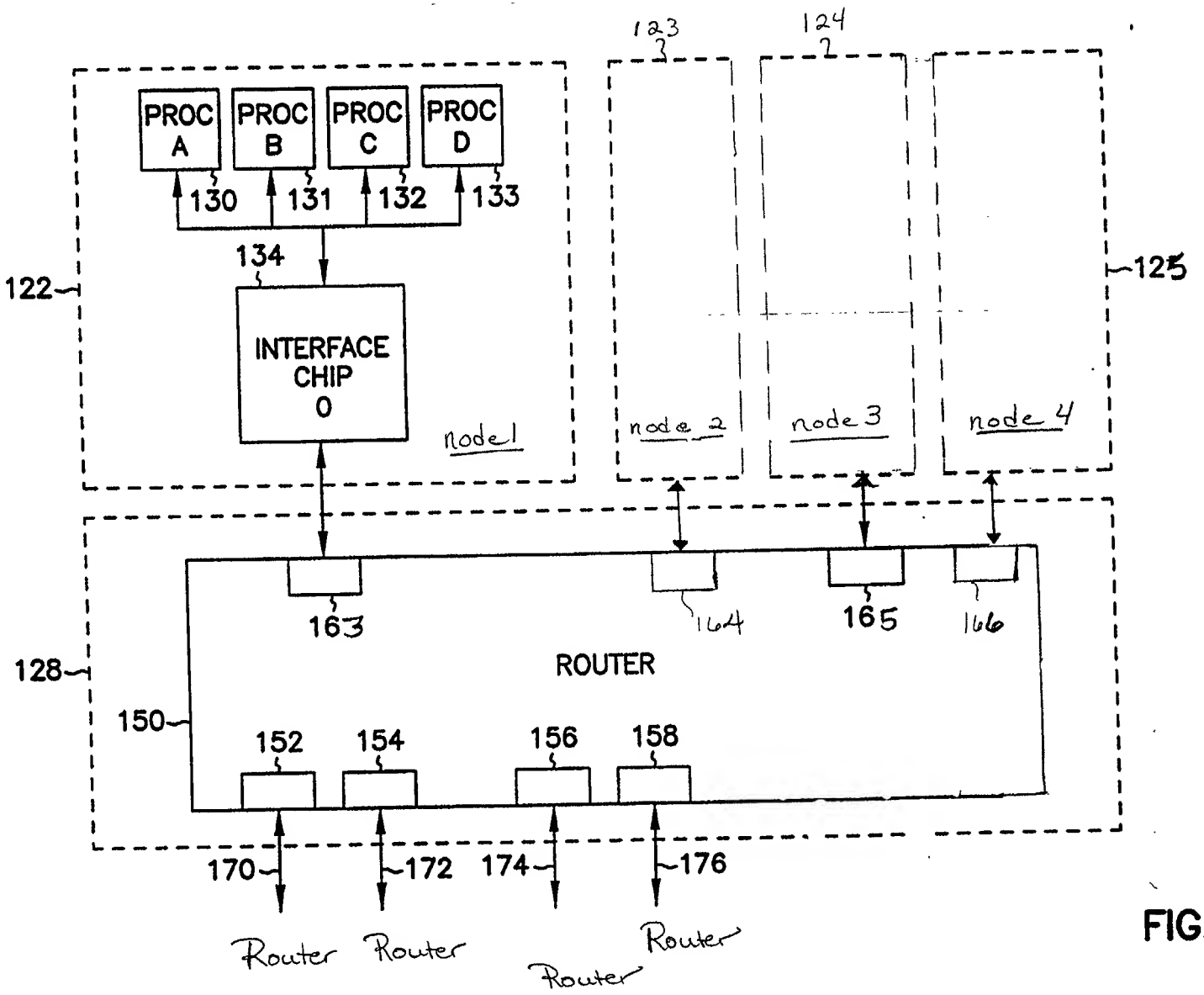

123
3
124
2
PROC A
PROC B
PROC C
PROC D
130
131
132
133
134
122
125
INTERFACE CHIP 0
node 1
node 2
node 3
node 4
163
164
165
166
128
ROUTER
150
152
154
156
158
170
172
174
176
Router
Router
Router
Router

FIG. 2

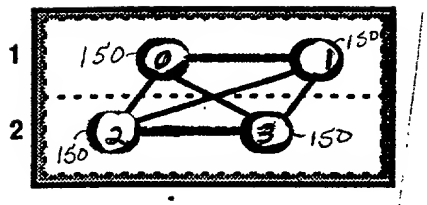
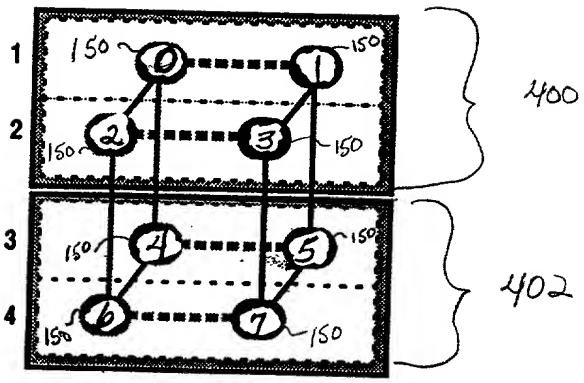


FIG. 3

FIG. 4



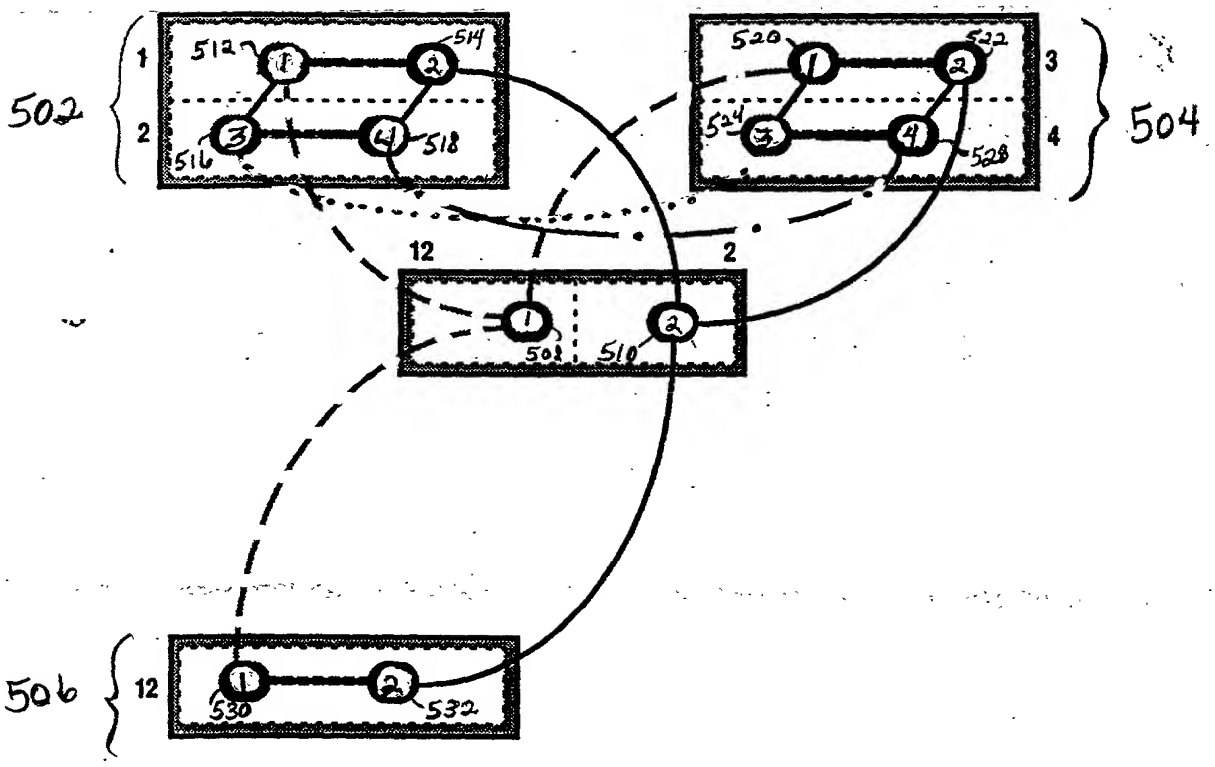

502
504
506
512
514
516
518
520
522
524
528
530
532
502
510
1
2
3
4
12

FIG. 5

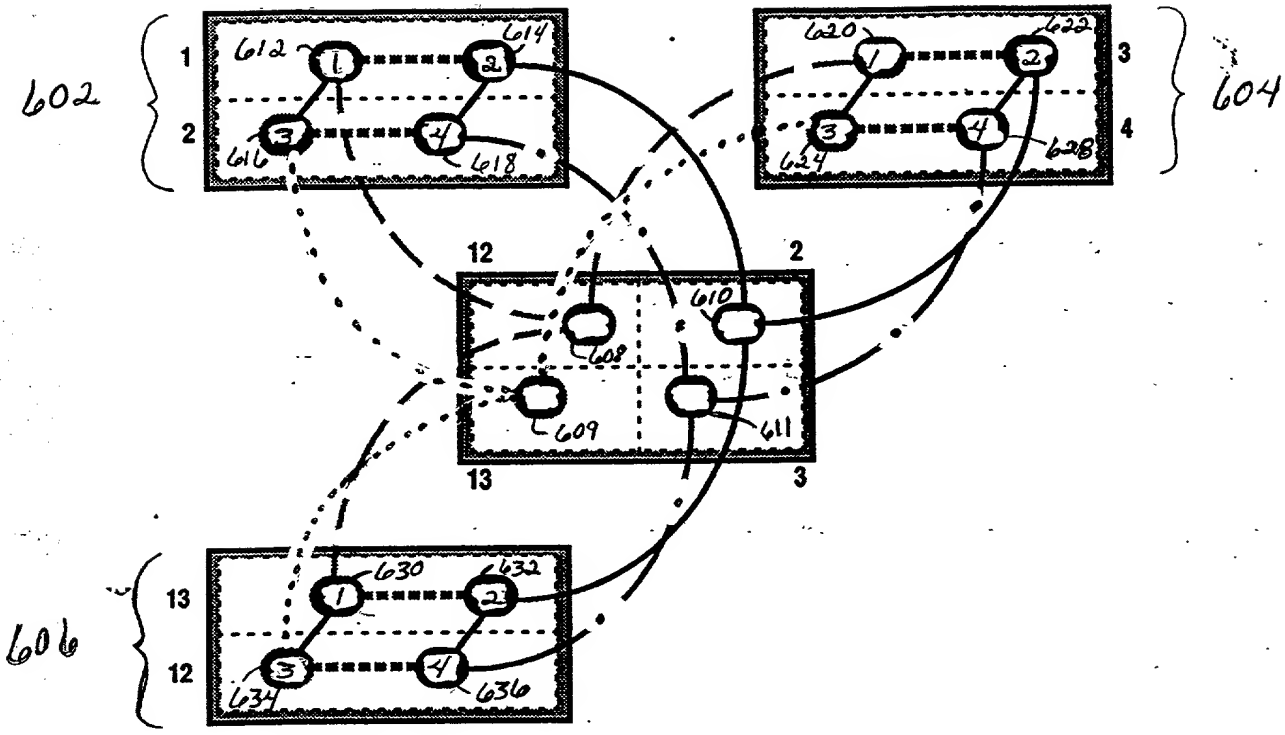

602
604
606
612
614
616
618
620
622
624
626
608
609
610
611
630
632
634
636
1
2
3
4
12
13

FIG. 6

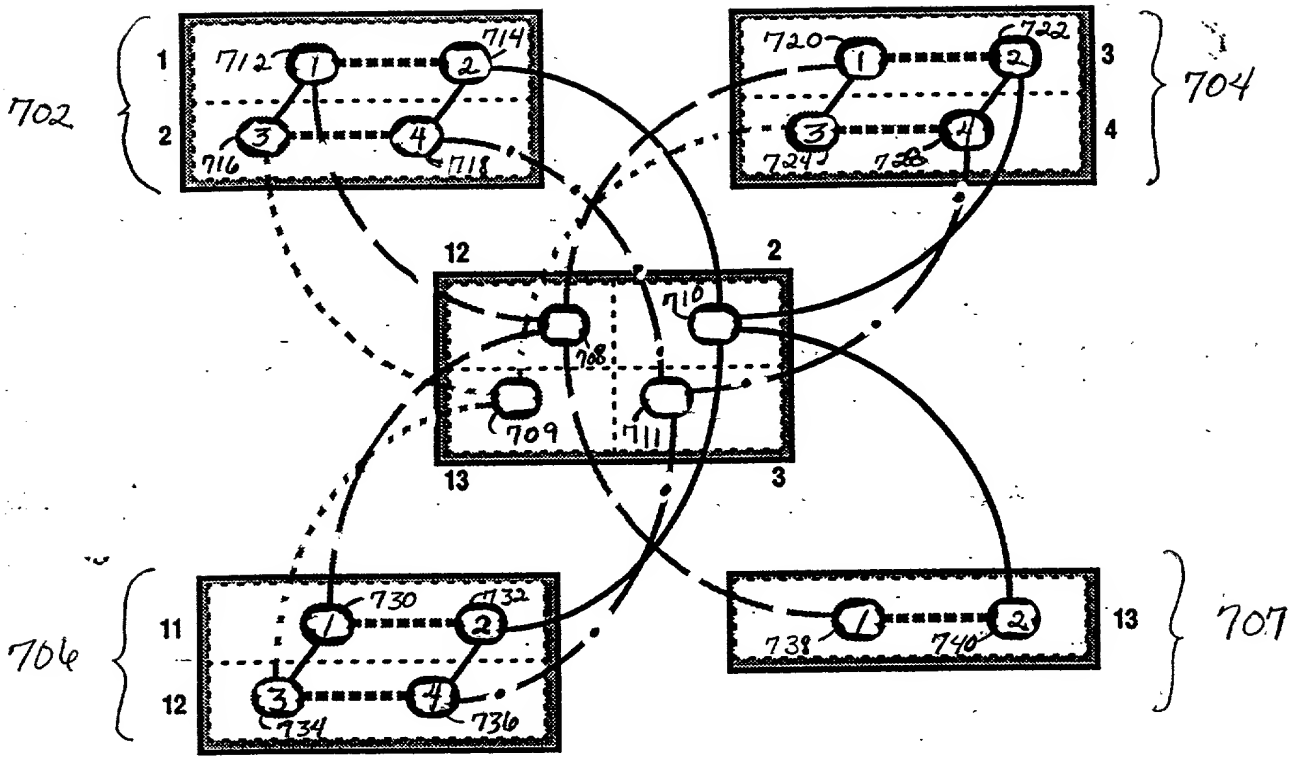

702
704
706
707
712
714
716
718
720
722
724
728
730
732
734
736
738
740
708
709
710
711
1
2
3
4
11
12
13

FIG. 7

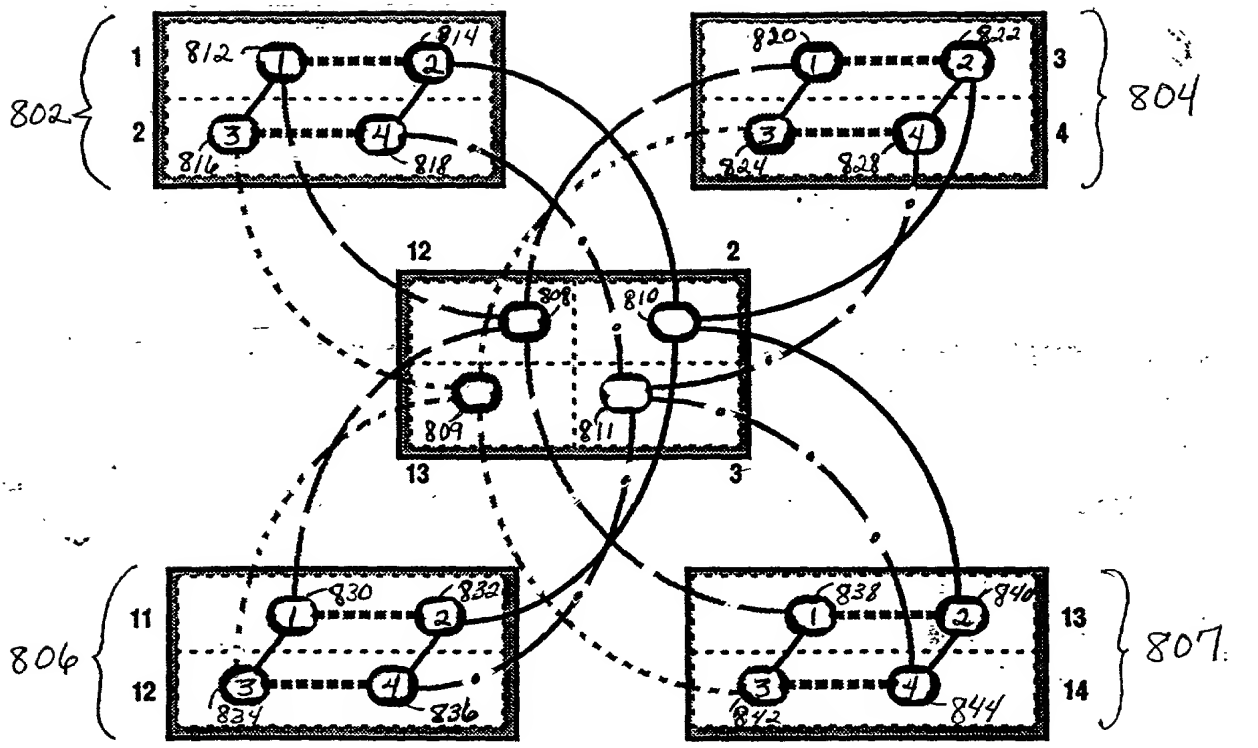

802
804
806
807
812
814
816
818
820
822
824
828
830
832
834
836
838
840
842
844
808
809
810
811
1
2
3
4
11
12
13
14

FIG. 8

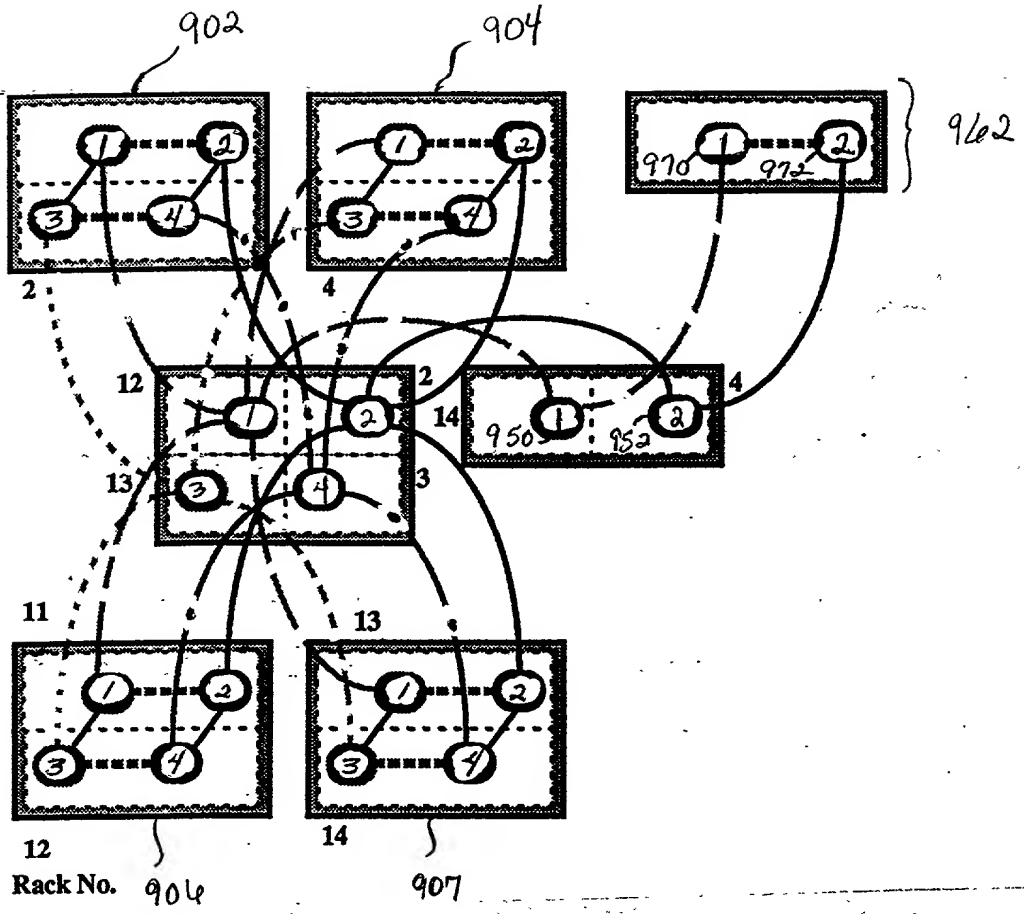

902
904
962
970
972
950
952
2
4
12
14
13
3
11
13
12
14
Rack No.
906
907

FIG. 9

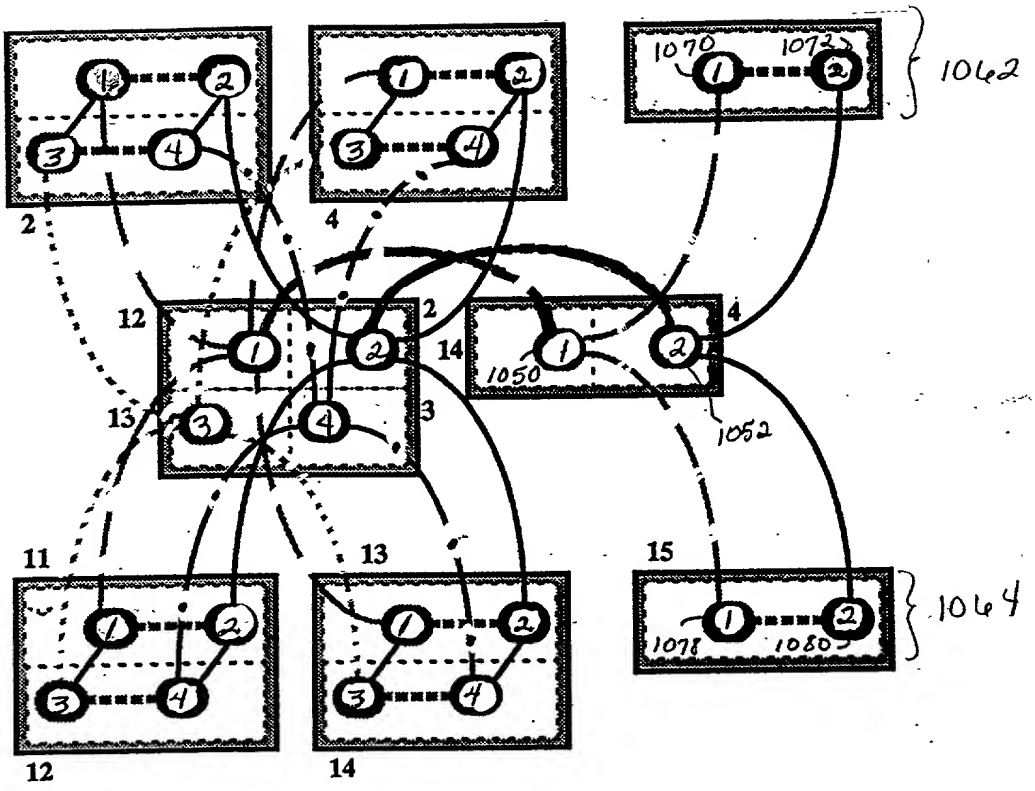

1070
1072
1062
2
4
12
2
14
4
1050
1052
13
3
11
13
15
1064
1078
1080
12
14
Rack No.

FIG. 10

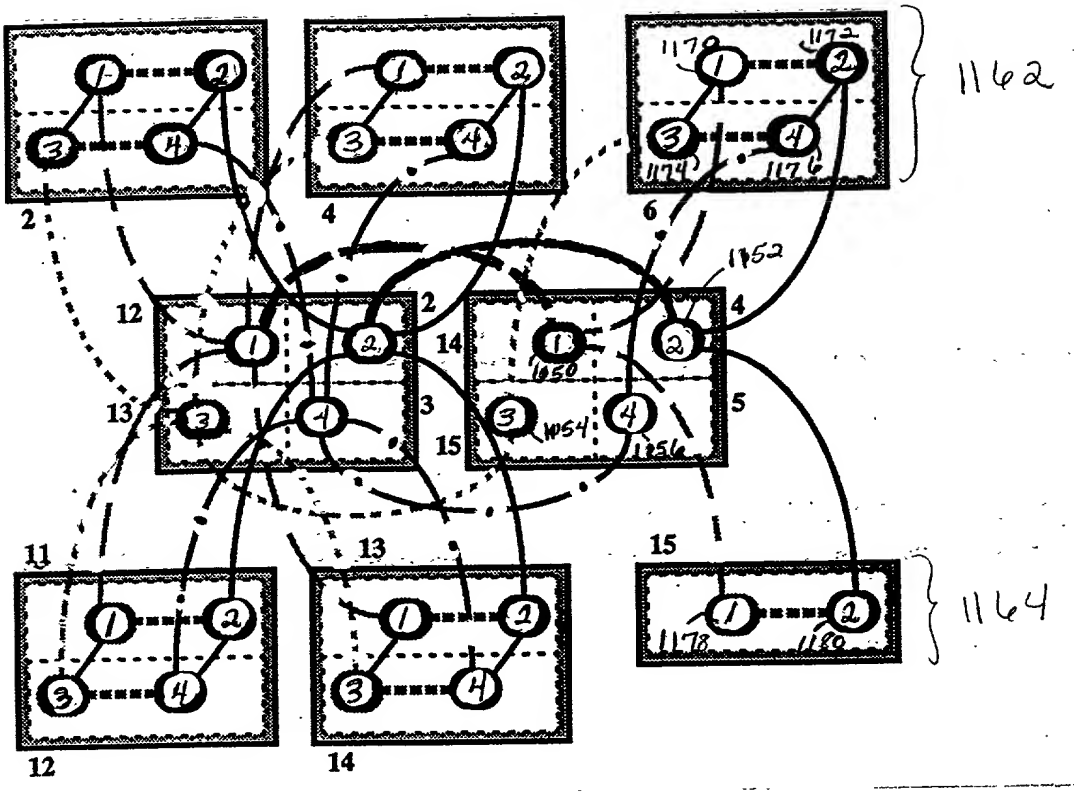

1162
1164
1152
1150
1154
1156
1170
1172
1174
1176
1178
1180
2
4
6
12
14
13
3
15
5
11
13
15
12
14

FIG. 11

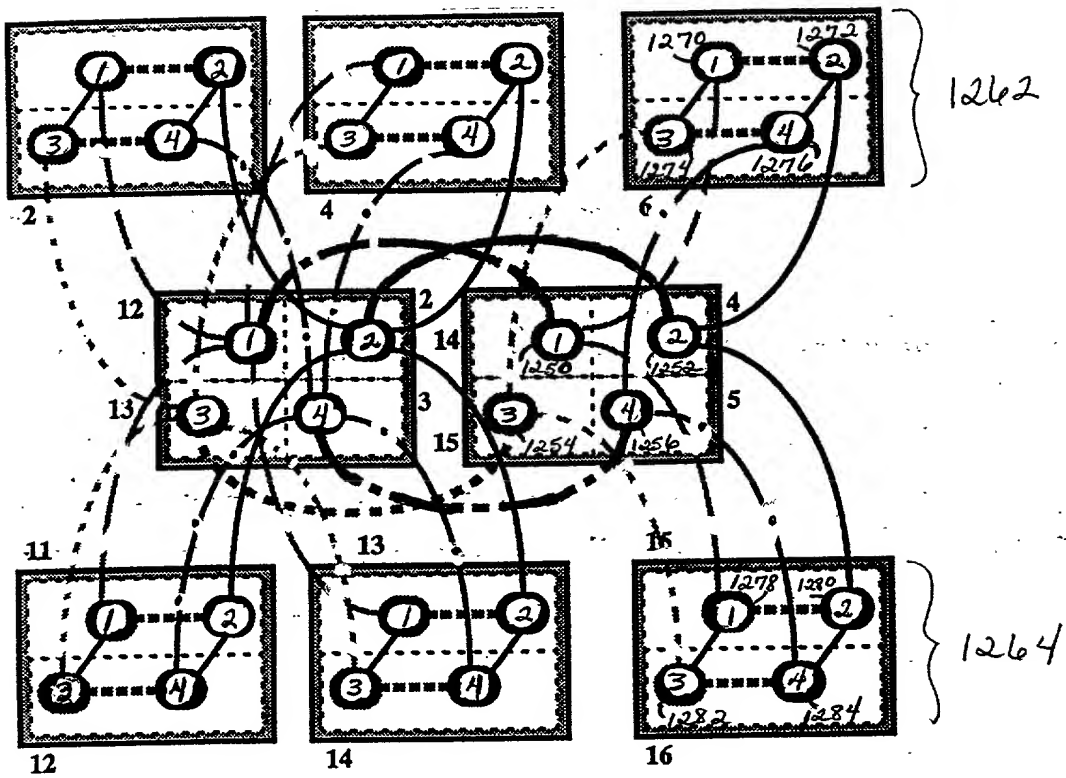

1262
1264
1270
1272
1274
1276
1250
1252
1254
1256
1278
1280
1282
1284
2
4
6
12
14
16
11
13
15

FIG. 12

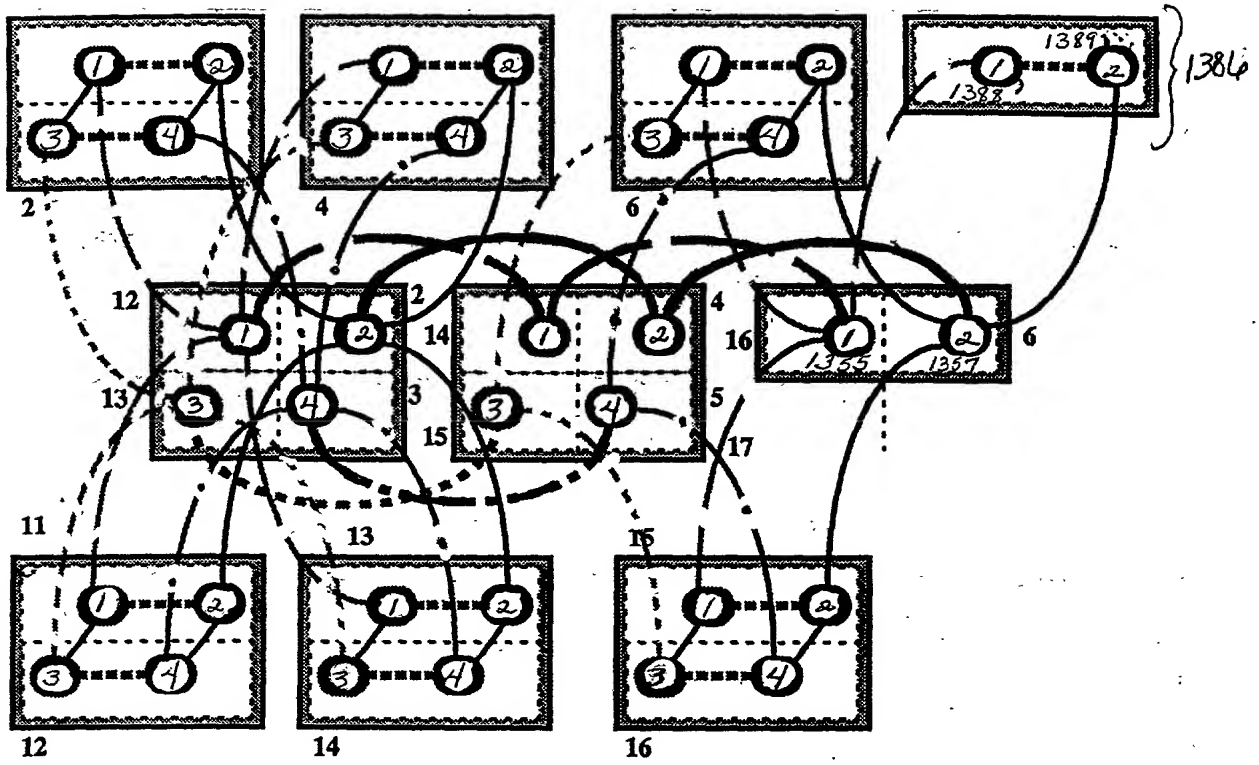

1389
1388
1386
2
4
6
12
2
14
4
16
6
13
3
15
5
17
11
13
15
12
14
16
Rack No.

FIG. 13

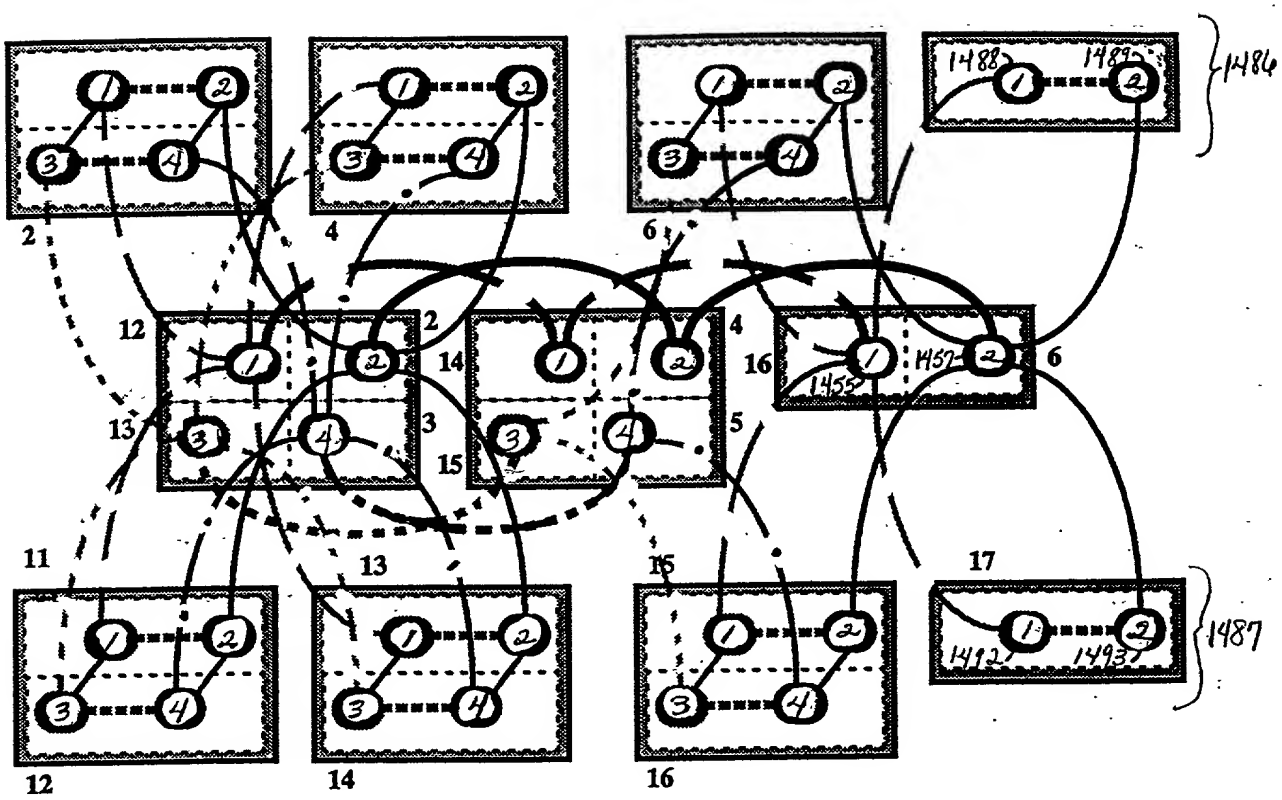


FIG. 14

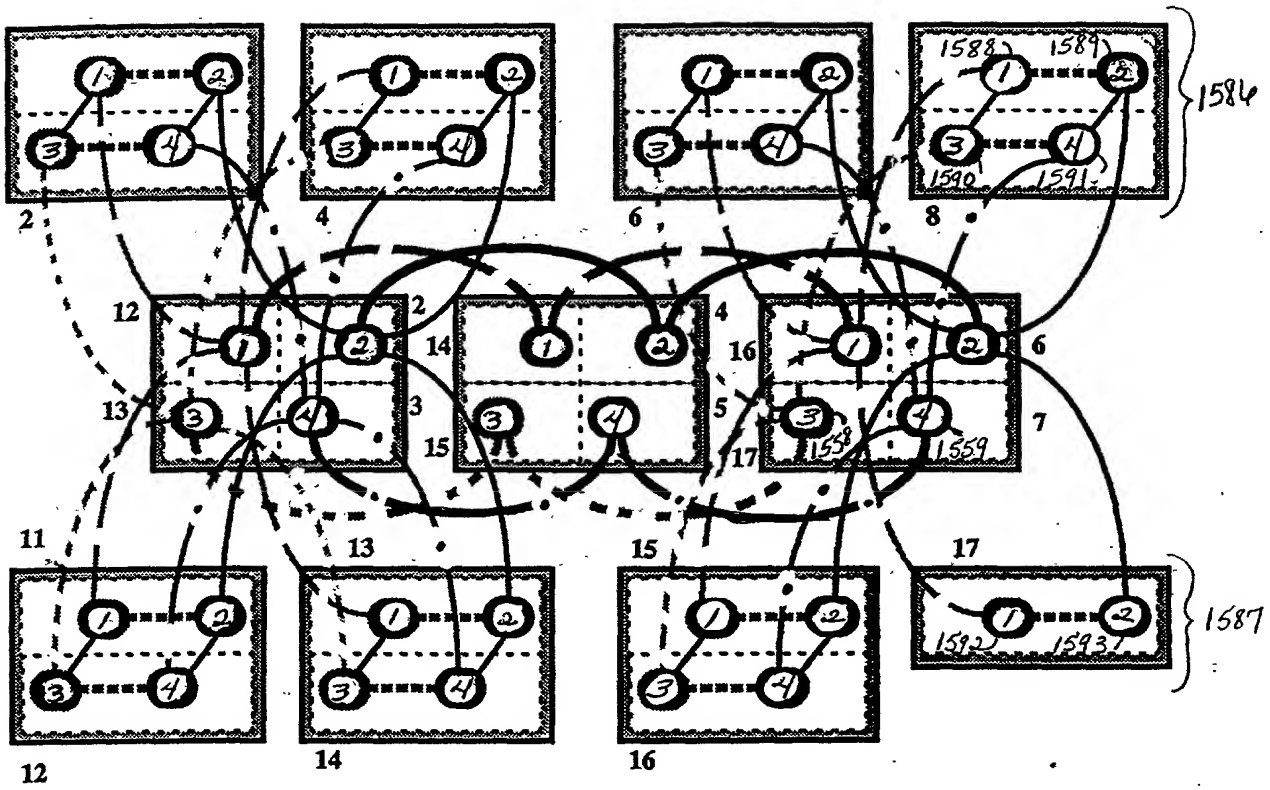

1588
1589
1584
1590
1591
2
4
6
8
12
2
14
4
16
6
13
3
5
7
15
17
1558
1559
11
13
15
17
1587
1592
1593
12
14
16
Rack No.

FIG. 15

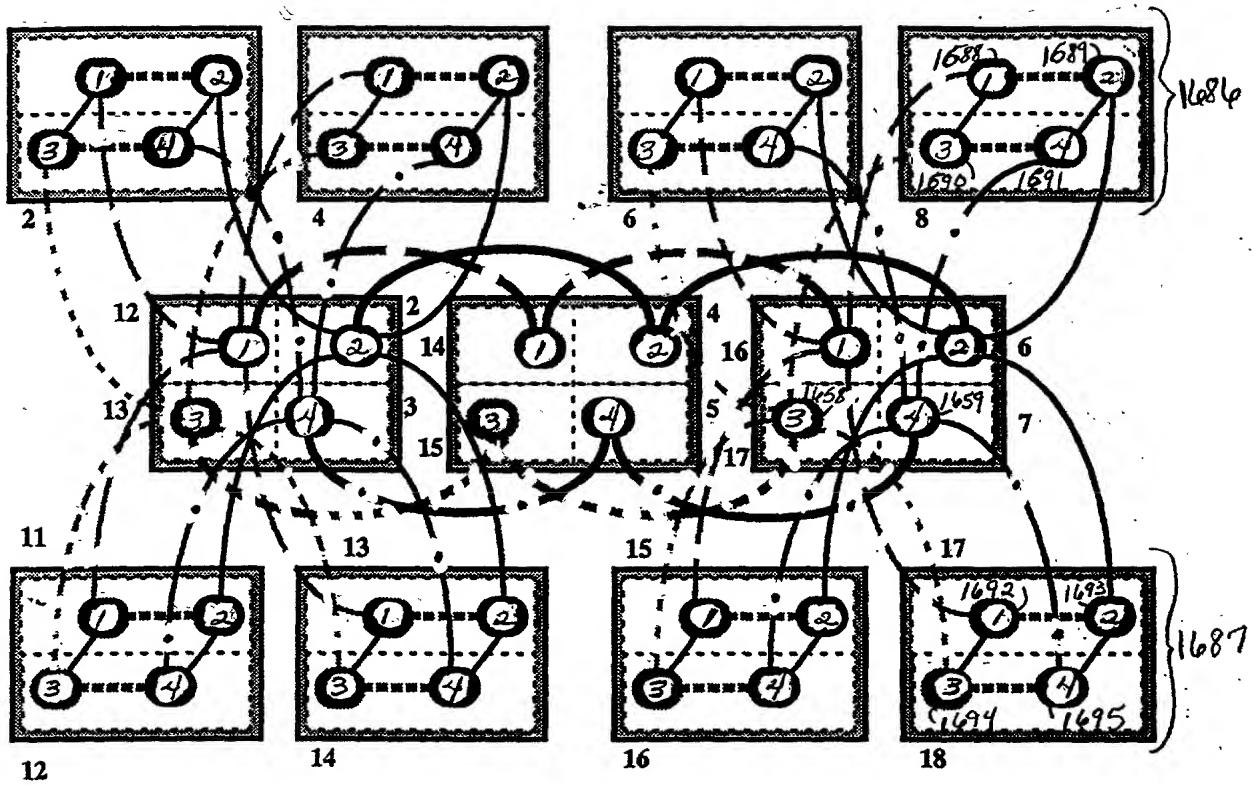

1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1658
1659
2
4
6
8
12
14
16
18
11
13
15
17
Rack No.

FIG. 16

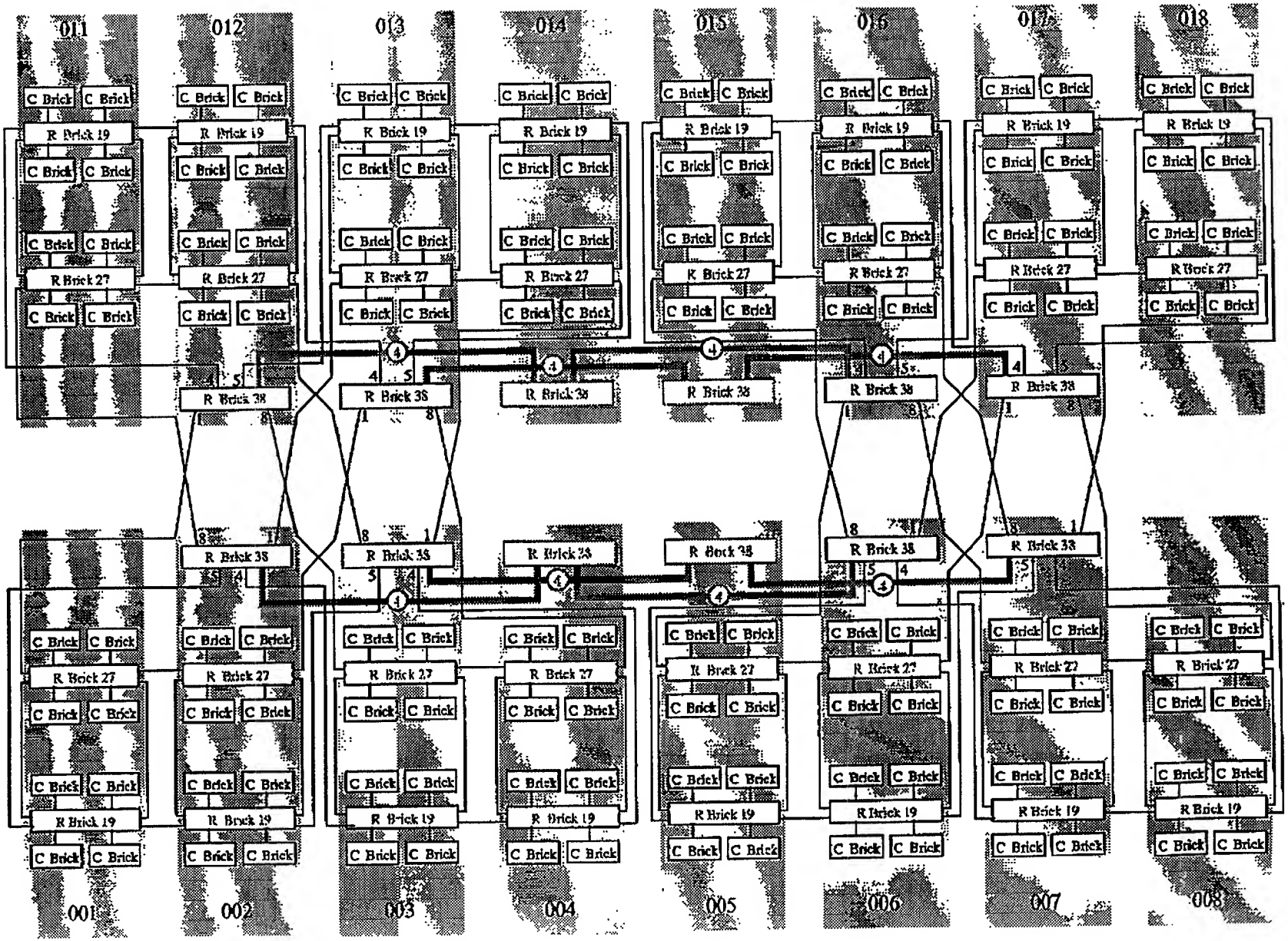

011
012
013
014
015
016
017
018
001
002
003
004
005
006
007
008
C Brick
R Brick 19
R Brick 27
R Brick 38

Fig. 17

SCHWEGMAN, LUNDBERG, WOESSNER & KLUTH, P.A.

United States Patent Application

COMBINED DECLARATION AND POWER OF ATTORNEY

As a below named inventor I hereby declare that: my residence, post office address and citizenship are as stated below next to my name; that

I verily believe I am the original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled: **NETWORK TOPOLOGY FOR A SCALABLE MULTIPROCESSOR SYSTEM**.

The specification of which is attached hereto.

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the patentability of this application in accordance with 37 C.F.R. § 1.56 (attached hereto). I also acknowledge my duty to disclose all information known to be material to patentability which became available between a filing date of a prior application and the national or PCT international filing date in the event this is a Continuation-In-Part application in accordance with 37 C.F.R. §1.63(e).

I hereby claim foreign priority benefits under 35 U.S.C. §119(a)-(d) or 365(b) of any foreign application(s) for patent or inventor's certificate, or 365(a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on the basis of which priority is claimed:

No such claim for priority is being made at this time.

I hereby claim the benefit under 35 U.S.C. § 119(e) of any United States provisional application(s) listed below:

No such claim for priority is being made at this time.

I hereby claim the benefit under 35 U.S.C. § 120 or 365(c) of any United States and PCT international application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT international application in the manner provided by the first paragraph of 35 U.S.C. § 112, I acknowledge the duty to disclose material information as defined in 37 C.F.R. § 1.56(a) which became available between the filing date of the prior application and the national or PCT international filing date of this application:

No such claim for priority is being made at this time.

I hereby appoint the following attorney(s) and/or patent agent(s) to prosecute this application and to transact all business in the Patent and Trademark Office connected herewith:

Adams, Gregory J.	Reg. No. P-44,494	Fordenbacher, Paul J.	Reg. No. 42,546	Nama, Kash	Reg. No. 44,255
Adams, Matthew W.	Reg. No. 43,459	Forrest, Bradley A.	Reg. No. 30,837	Nelson, Albin J.	Reg. No. 28,650
Anglin, J. Michael	Reg. No. 24,916	Harris, Robert J.	Reg. No. 37,346	Nielsen, Walter W.	Reg. No. 25,539
Arora, Suneel	Reg. No. 42,267	Huebsch, Joseph C.	Reg. No. 42,673	Oh, Allen J.	Reg. No. 42,047
Bianchi, Timothy E.	Reg. No. 39,610	Jurkovich, Patti J.	Reg. No. P-44,813	Padys, Danny J.	Reg. No. 35,635
Billion, Richard E.	Reg. No. 32,836	Kalis, Janal M.	Reg. No. 37,650	Parker, J. Kevin	Reg. No. 33,024
Black, David W.	Reg. No. 42,331	Kaufmann, John D.	Reg. No. 24,017	Peacock, Gregg A.	Reg. No. P-45,001
Brennan, Leoniede M.	Reg. No. 35,832	Klima-Silberg, Catherine I.	Reg. No. 40,052	Perdok, Monique M.	Reg. No. 42,989
Brennan, Thomas F.	Reg. No. 35,075	Kluth, Daniel J.	Reg. No. 32,146	Polglaze, Daniel J.	Reg. No. 39,801
Brooks, Edward J., III	Reg. No. 40,925	Lacy, Rodney L.	Reg. No. 41,136	Prout, William F.	Reg. No. 33,995
Byrne, Christopher	Reg. No. 32,204	Leffert, Thomas W.	Reg. No. 40,697	Schwegman, Micheal L.	Reg. No. 25,816
Chu, Dinh C.P.	Reg. No. 41,676	Lemaire, Charles A.	Reg. No. 36,198	Sieffert, Kent J.	Reg. No. 41,312
Clark, Barbara J.	Reg. No. 38,107	Litman, Mark A.	Reg. No. 26,390	Slifer, Russell D.	Reg. No. 39,838
Dahl, John M.	Reg. No. P-44,639	Lundberg, Steven W.	Reg. No. 30,568	Steffey, Charles E.	Reg. No. 25,179
Drake, Eduardo E.	Reg. No. 40,594	Mack, Lisa K.	Reg. No. 42,825	Terry, Kathleen R.	Reg. No. 31,884
Eliseeva, Maria M.	Reg. No. 43,328	Maki, Peter C.	Reg. No. 42,832	Viksnins, Ann S.	Reg. No. 37,748
Embretson, Janet E.	Reg. No. 39,665	Malen, Peter L.	Reg. No. P-44,894	Weiner, Steve	Reg. No. 38,330
Fernandez, Irene	Reg. No. 34,625	Mates, Robert E.	Reg. No. 35,271	Woessner, Warren D.	Reg. No. 30,440
Fogg, David N.	Reg. No. 35,138	McCrackin, Ann M.	Reg. No. 42,858		

I hereby authorize them to act and rely on instructions from and communicate directly with the person/assignee/attorney/firm/organization/who/which first sends/sent this case to them and by whom/which I hereby declare that I have consented after full disclosure to be represented unless/until I instruct Schwegman, Lundberg, Woessner & Kluth, P.A. to the contrary.

Please direct all correspondence in this case to **Schwegman, Lundberg, Woessner & Kluth, P.A.** at the address indicated below:

P.O. Box 2938, Minneapolis, MN 55402
Telephone No. (612)373-6900

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of joint inventor number 1 : **Martin M. Deneroff**
Citizenship: **United States of America** Residence: **Palo Alto, CA**
Post Office Address: 2970 South Court Street
Palo Alto, CA 94306

Signature: _____ Date: _____
Martin M. Deneroff

Full Name of joint inventor number 2 : **Gregory M. Thorson**
Citizenship: **United States of America** Residence: **Altoona, WI**
Post Office Address: 1119 Sweet Water Close
Altoona, WI 54720

Signature: _____ Date: _____
Gregory M. Thorson

X Additional inventors are being named on separately numbered sheets, attached hereto.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of joint inventor number 3 : **Randal S. Passint**
Citizenship: **United States of America** Residence: **Chippewa Falls, WI**
Post Office Address: 9550 167 Street
Chippewa Falls, WI 54729

Signature: _____ Date: _____
Randal S. Passint

Full Name of inventor:
Citizenship: Residence:
Post Office Address:

Signature: _____ Date: _____

Full Name of inventor:
Citizenship: Residence:
Post Office Address:

Signature: _____ Date: _____

Full Name of inventor:
Citizenship: Residence:
Post Office Address:

Signature: _____ Date: _____

§ 1.56 Duty to disclose information material to patentability.

(a) A patent by its very nature is affected with a public interest. The public interest is best served, and the most effective patent examination occurs when, at the time an application is being examined, the Office is aware of and evaluates the teachings of all information material to patentability. Each individual associated with the filing and prosecution of a patent application has a duty of candor and good faith in dealing with the Office, which includes a duty to disclose to the Office all information known to that individual to be material to patentability as defined in this section. The duty to disclose information exists with respect to each pending claim until the claim is canceled or withdrawn from consideration, or the application becomes abandoned. Information material to the patentability of a claim that is canceled or withdrawn from consideration need not be submitted if the information is not material to the patentability of any claim remaining under consideration in the application. There is no duty to submit information which is not material to the patentability of any existing claim. The duty to disclose all information known to be material to patentability is deemed to be satisfied if all information known to be material to patentability of any claim issued in a patent was cited by the Office or submitted to the Office in the manner prescribed by §§ 1.97(b)-(d) and 1.98. However, no patent will be granted on an application in connection with which fraud on the Office was practiced or attempted or the duty of disclosure was violated through bad faith or intentional misconduct. The Office encourages applicants to carefully examine:

(1) prior art cited in search reports of a foreign patent office in a counterpart application, and

(2) the closest information over which individuals associated with the filing or prosecution of a patent application believe any pending claim patentably defines, to make sure that any material information contained therein is disclosed to the Office.

(b) Under this section, information is material to patentability when it is not cumulative to information already of record or being made of record in the application, and

(1) It establishes, by itself or in combination with other information, a prima facie case of unpatentability of a claim; or

(2) It refutes, or is inconsistent with, a position the applicant takes in:

(i) Opposing an argument of unpatentability relied on by the Office, or

(ii) Asserting an argument of patentability.

A prima facie case of unpatentability is established when the information compels a conclusion that a claim is unpatentable under the preponderance of evidence, burden-of-proof standard, giving each term in the claim its broadest reasonable construction consistent with the specification, and before any consideration is given to evidence which may be submitted in an attempt to establish a contrary conclusion of patentability.

(c) Individuals associated with the filing or prosecution of a patent application within the meaning of this section are:

(1) Each inventor named in the application:

(2) Each attorney or agent who prepares or prosecutes the application; and

(3) Every other person who is substantively involved in the preparation or prosecution of the application and who is associated with the inventor, with the assignee or with anyone to whom there is an obligation to assign the application.

(d) Individuals other than the attorney, agent or inventor may comply with this section by disclosing information to the attorney, agent, or inventor.